# **Factor Analysis on Ozone Level Detection**

## Introduction

Ozone day is the day when weather condition mix with the pollution of high-level ozone close to the ground. Ground level ozone(O3) is not created directly but formed as the result of complex chemical reaction called Volatile Organic Compounds. Short-term exposure to elevated ambient ozone not only can cause a number of health problems for human but also harmful to corpses (Fan & Wei, 2008).

Texas Commission on Environmental Quality is working on how to build accurate alarm forecasting model for the Houston area (Fan & Wei, 2008). However, there are numerous challenges to construct the model such as sparse and evolving dataset, limited data size, large irrelevant features, and the true model is stochastic as a function of measurable factors. The focus of this paper is to choose the most reliable factors that determine the future of the ozone level days with unknown parameter and different distribution. 15 statistical techniques applied in 2003 (Schlink, et al., 2003) and the results are none of them perform better from one to others in all aspects.

The author of this analysis realizes that the model that will be conducted in the future cannot fully predict ground-level ozone accurately due to limited knowledge of the author with the highly skewed dataset. However, the author will try to predict what are the main factors that contribute to the higher probability of the ozone level.

## Dataset Explanation

The dataset is taken from <https://archive.ics.uci.edu/ml/datasets/ozone+level+detection>. This analysis will use one-hour peak dataset with the file name *ohehr.data*. There are 10 categories in this dataset and they are measured in different level. The total variables in this dataset are 72 variables. Here are the table describing the categories and variables.

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Label** | **Information** | **Measured at** |
| Temperature | T |  | 22 times and three levels |
| Wind speed rate | wsr |  | 22 times and three levels |
| Relative humidity | rh | percentage of partial pressure water vapor to the equilibrium vapor pressure | 3 levels |
| East-West direction wind | u |  | 3 levels |
| North-South direction Wind | v |  | 3 levels |
| Geopotential height | ht | vertical coordinate referenced to Earth's mean sea level | 3 levels |
| K Index | ki | quantification of the disturbances in the horizontal component of earth’s magnetic field. >5 is geomagnetic storm |  |
| T Total | tt | Index to assess storm strength. <44 less likely and >56 scattered severed storms |  |
| Sea level Pressure | slp | the pressure within the atmosphere of earth | current and previous day |
| Attribute | attribute | 0 when normal day and 1 when ozone day |  |

The three levels in the dataset reflect to the difference in atmospheric pressure which is 500,700 and 850 Hectopascal or roughly 5500, 3100 and 1500 meters height respectively

This analysis will start by assessing univariate and multivariate normality, conducting principal component analysis, factor analysis ,clustering and two mean sample hoteling T2 test. Since There are 1848 observations and 72 variables in this dataset, the author decided only to include 25 variables and omit 47 variables which are temperature and windspeed rate at 23 different times.

# METHODOLIGAL ANALYSIS

## Descriptive Statistics

Based on the boxplot of temperature and wind speed rate (Figure 1 and Figure 2) at 22 different times, the only information that can be extracted is that as the temperature increase, the windspeed will also increase. Based on the boxplot of five variables (rh,u,v, ht,t) on Figure 3, temperature, relative humidity and north-south direction wind(v) are negatively correlated with the atmospheric meanwhile east-west wind direction (u), and geopotential height is positively correlated with the atmospheric pressure.

## Assessment of Normality

Based on Shapiro Wilk test of each variable(Figure 5), there is only one variable that contains partial univariate normality which is north-south direction wind at 850 hpa with p-value 0.3. Therefore, the author uses boxcox transformation to find the optimum lambda to make the distribution more normal. In result, most of the lambda doesn’t give even partial normality (Figure 4) to the variable. Peak windspeed rate will have cube square root transformation because lambda is close to 0.333

The dataset does not pass univariate normality and it will result that the dataset is not multivariate normally distributed. The result of Mardia test, Chi-square plot to the squared Mahalanobis distance (Figure 6, Figure 7, and Figure 8) and the proportion of the contour give the same result. Therefore, the dataset is not normally distributed.

## Correlation Matrix

The correlation matrix is quite hard to interpret but several takeouts from this standardized correlation matrix are geopotential height in general positively correlated with temperature meanwhile temperature, and geopotential height are negatively correlated with east-west wind direction. Relative humidity is positively correlated with the index to storm strength. However, correlation does not mean causation. Therefore, the author will not conclude anything based on this analysis.

## Principal Component Analysis

Based on the scree plot of principal component analysis(Figure 10), the first component can explain 32.4% total variation in the dataset followed by 19.6%, 9%, and 7% for second, third and fourth factor respectively. Total variation explained by four components is 68%. Because adding more variable does not add significant total variance, this analysis will only use four principal components for further analysis. Because the sum of squared eigenvectors in each dimension is one. It means that squared eigenvector of each variable is the contribution of given variable to given component.

Based on Figure 11, Temperature and geopotential height are the most critical component for component one and those variables will be the contrast of east-west wind direction in the same component. The second component will be dominantly by North-South wind direction and contrast with sea level pressure. This analysis is in general form, and factor analysis will have a deeper analysis of the similarity and dissimilarity among variables.

## Factor Analysis

Figure 12 is the bar chart of factor loadings of each variables using varimax rotation. The red bar reflects positive loadings, and blue bar reflects negative loadings. The noticeable variables that contribute to shape factor one is the contrast between temperature and geopotential height with east wind direction. This factor will be called temperature and geopotential height. Furthermore, component two are mostly windspeed rate and north-south direction. Factor two will be called windspeed to north-south.

The author creates the group of the variables based on the pressure level (500,700, and 850 Hpa) and turns out that the higher the pressure, the more this group contributes to Factor 1 and Factor 2 (Figure 13). It seems that there is more information that can be extracted at higher pressure level. Furthermore, temperature plays important role in both factors and based on the journal (Fan & Wei, 2008), more ozone days occurred in high temperature compared to low temperature. However, this information is not enough to explain why ozone occurred.

Based on the squared factor’s loading, the first two variables that contribute the most to the Factor 1 are K-index and t-total followed by temperature and sea level pressure.

## Two sample Variance Covariance Matrices

Based on the Box M of equal variances, the group with attribute 0 and 1 do not have equal variance covariance matrix. The Chi-square approximation is 598,36 with 300 degrees of freedom (Figure 18).The two sample T2 test shows the test statistic of 794 and follows chi-square distribution with 24 degree of freedom (Figure 20). Therefore, the multivariate mean group for ozone and non-ozone day are not equal to zero.

Based on the Figure 19, individual test of the ANOVA shows that relative humidity, k-index, t-total, sea level pressure and precipitation are not different between the group ozone day and non-ozone day. It seems that the windspeed, temperature and direction of the wind is different among the group. The authors conduct test of MANOVA on three different level of pressure and turns out that altogether (v,u,t,rh and ht) are different between ozone day and non-ozone day.

## Conclusion

Appendix

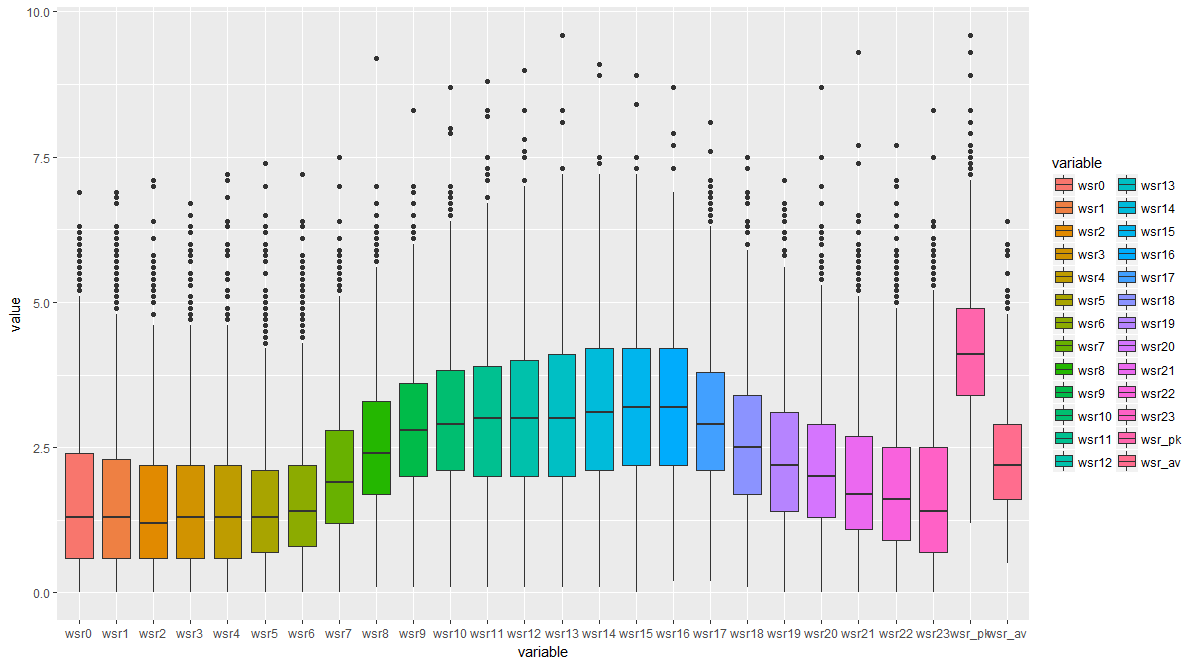


Figure Boxplot of temperature at 22 different times

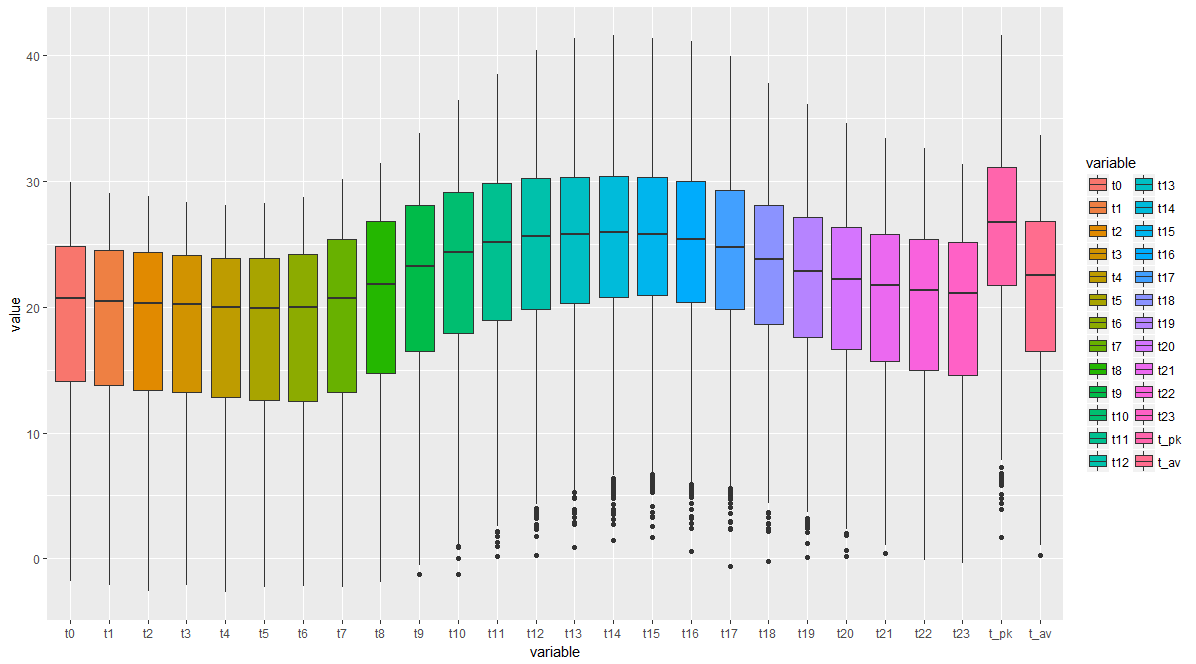


Figure Boxplot of temperature at 22 different times.

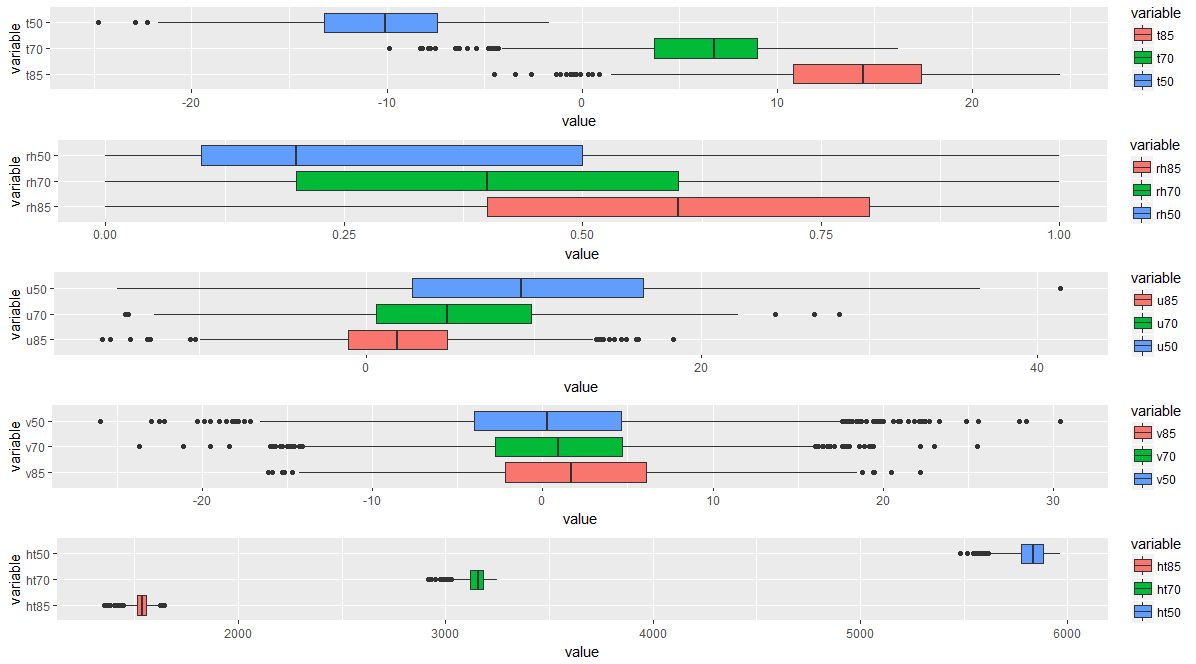


Figure Boxplot of 5 variables (t,rh,u,v,ht) at three different levels.

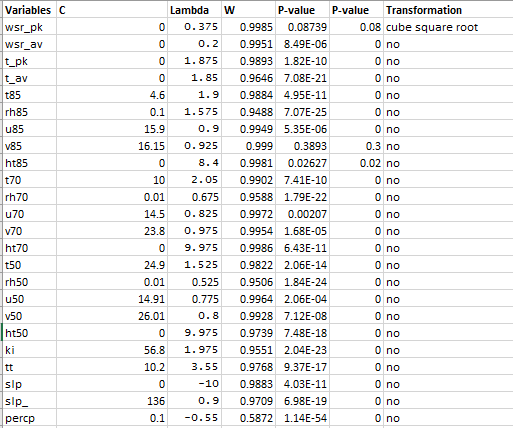
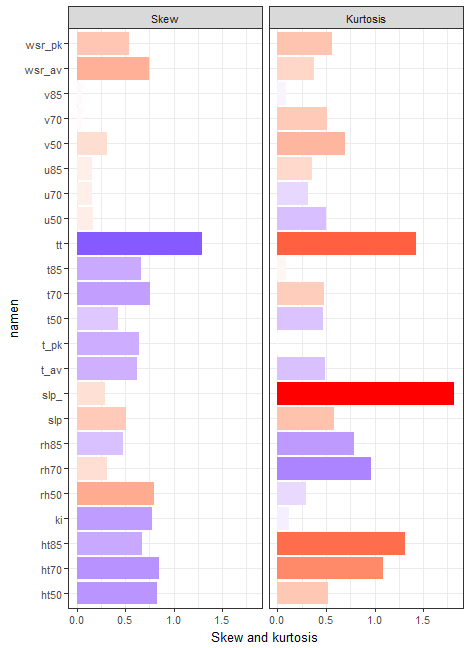


Figure Bar chart of skewness and kurtosis

Figure Table of box cox transformation

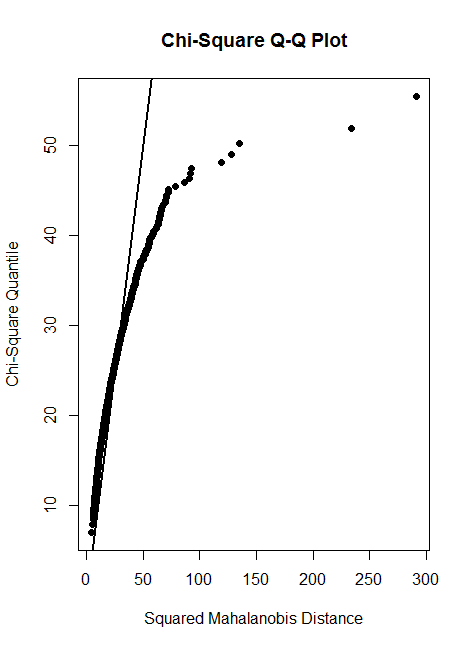
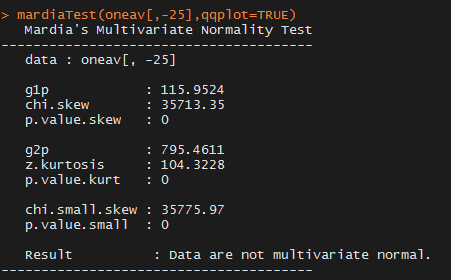


Figure Output of Mardia Test

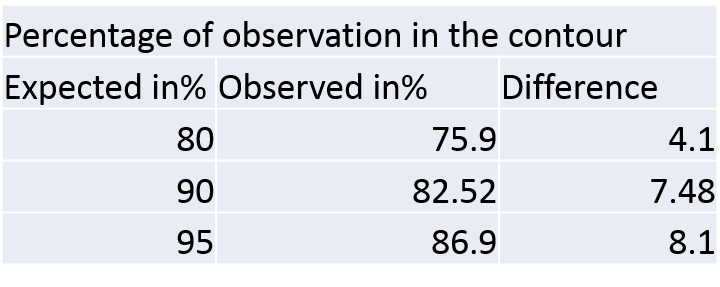


Figure Plot of Chi Square quantile and Squared Mahalanobis Distance

Figure Multivariate normality expected and observed proportion

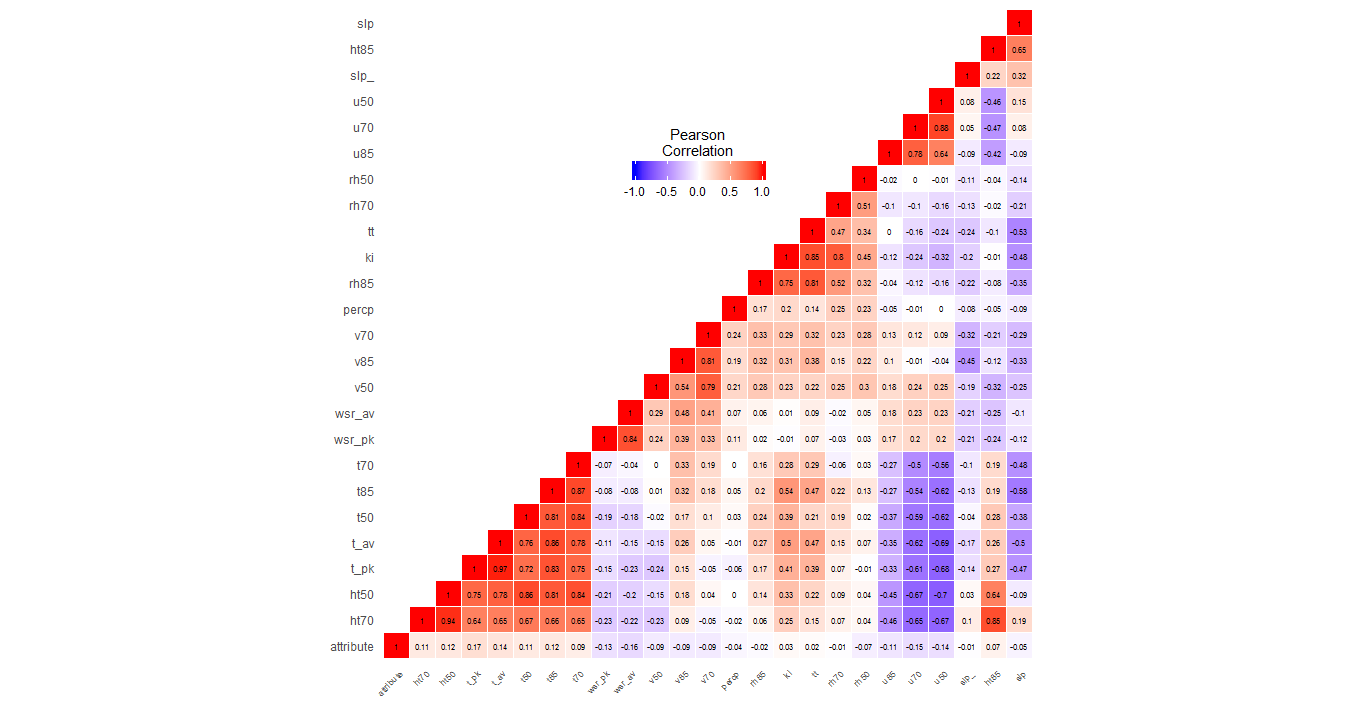


Figure Correaltion Matrix

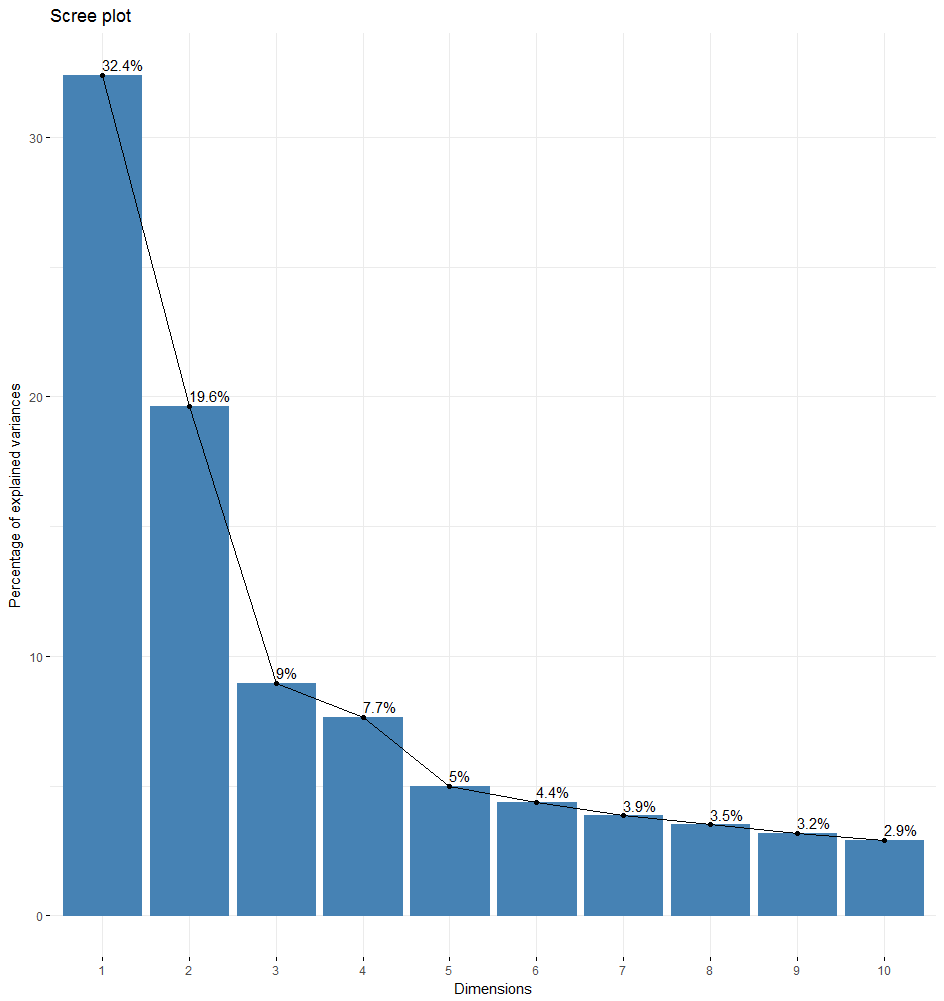


Figure Scree Plot of Principal Component Analysis

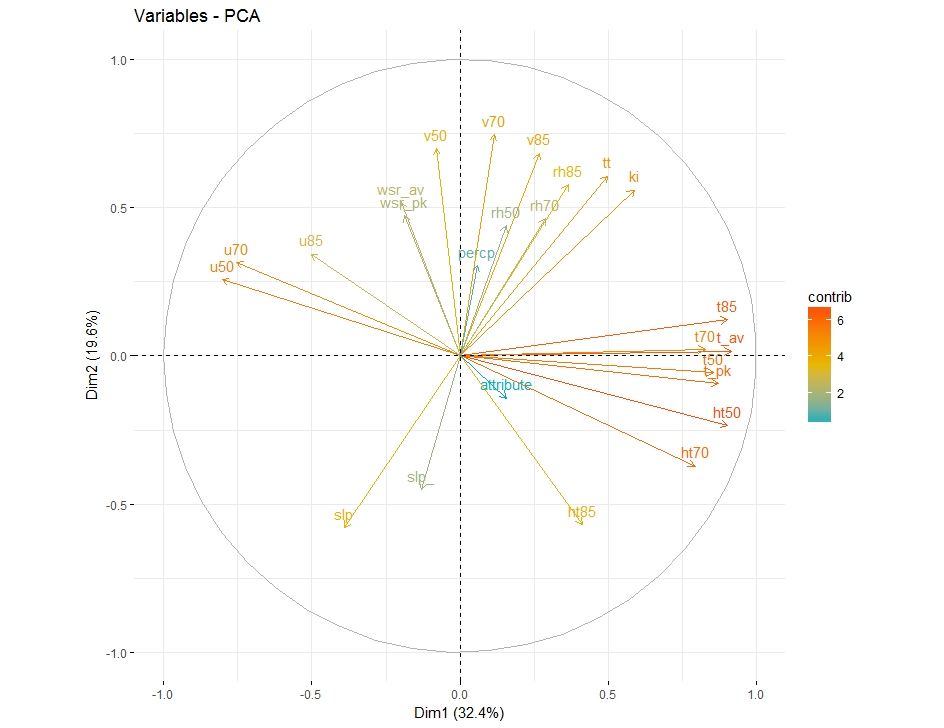


Figure Direction and coordinate of each variable in the first two principal components

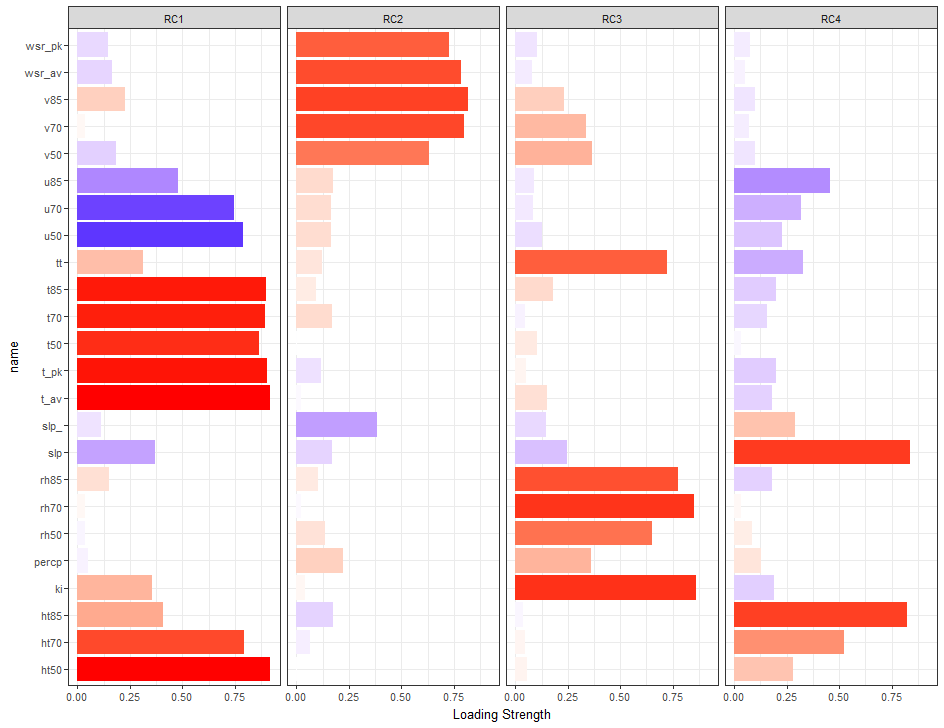


Figure factor loadings using varimax rotation

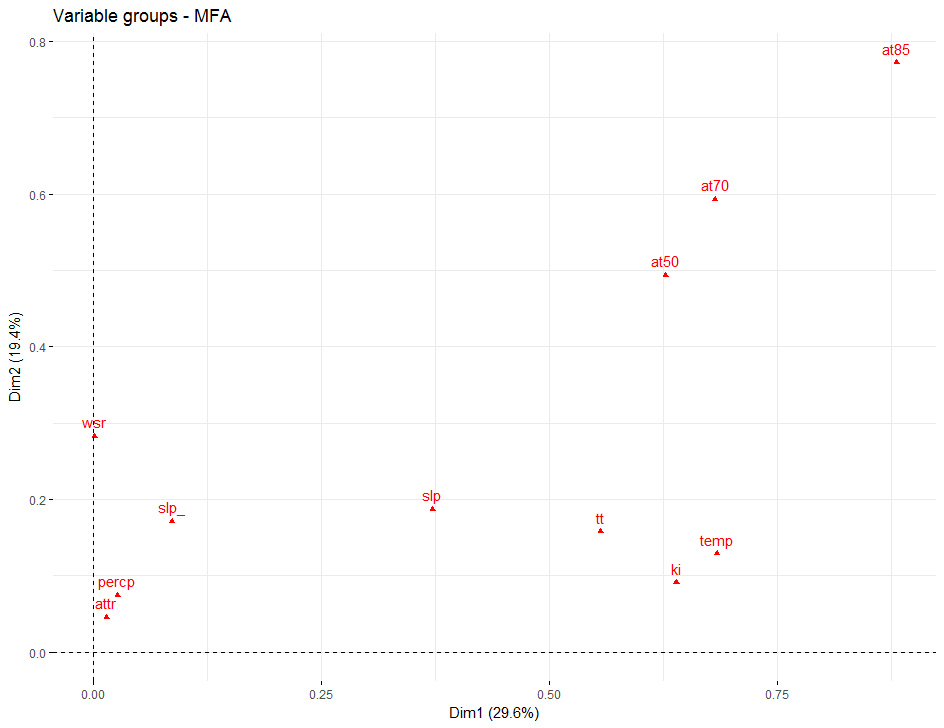


Figure Plot of grouping at different level of pressure

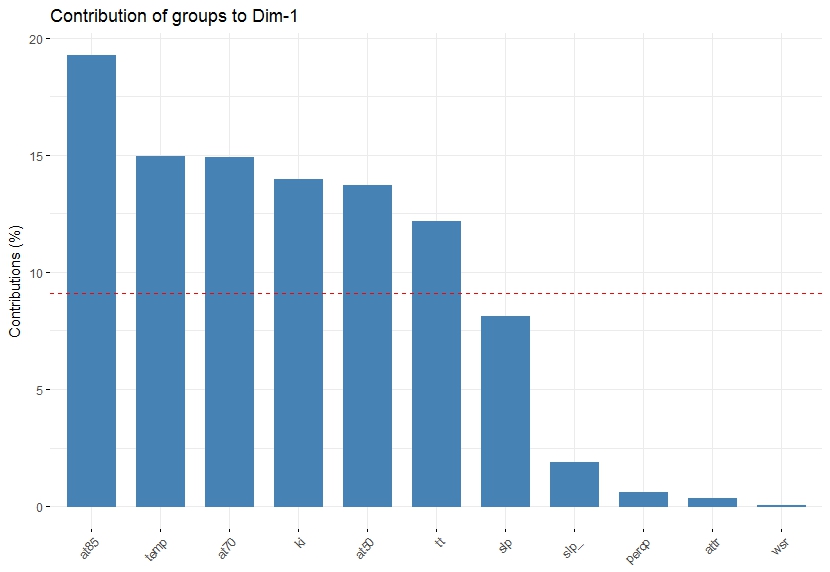


Figure Contribution of the group to dimension 1

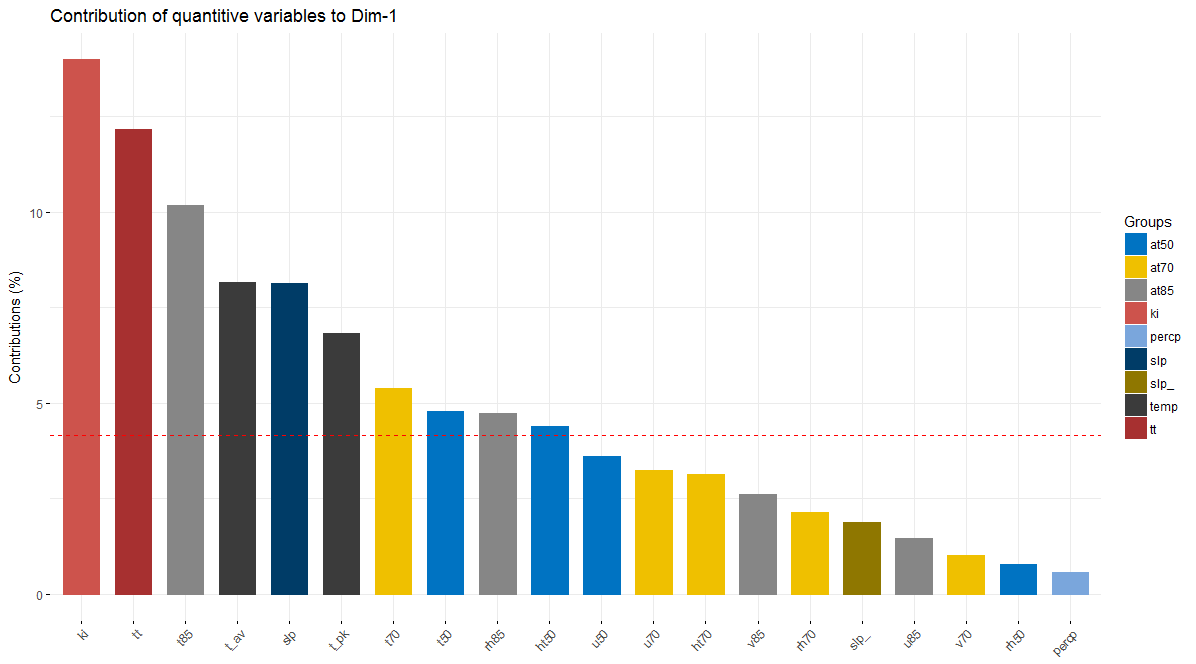


Figure Contribution of each variable to Factor 1

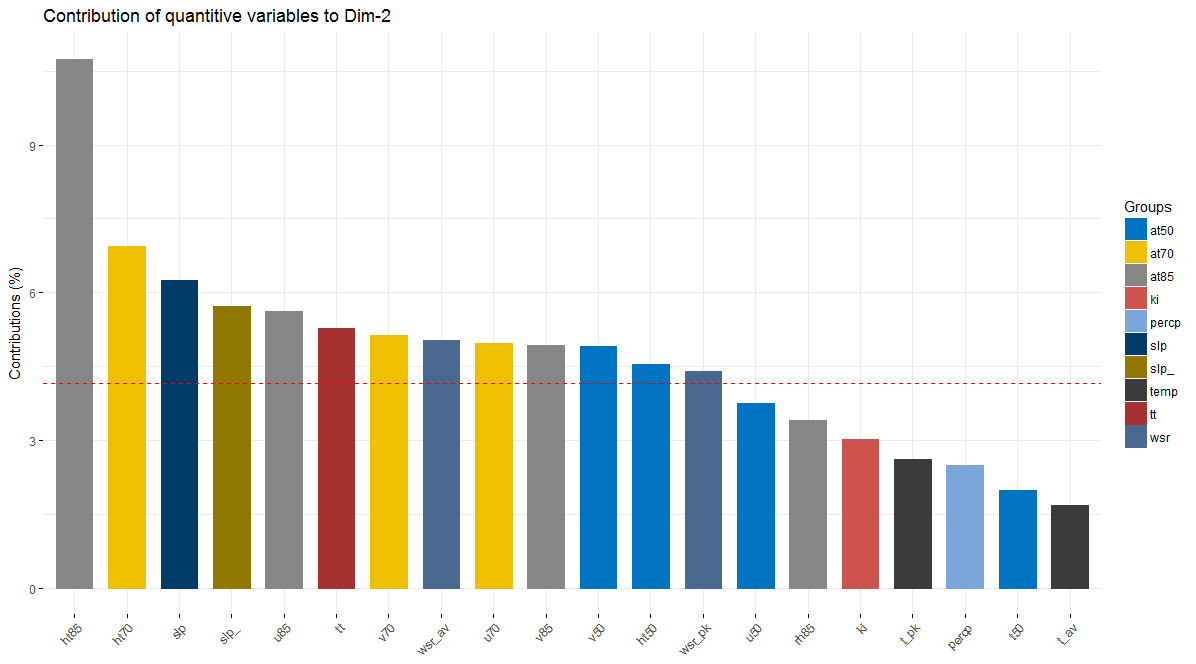
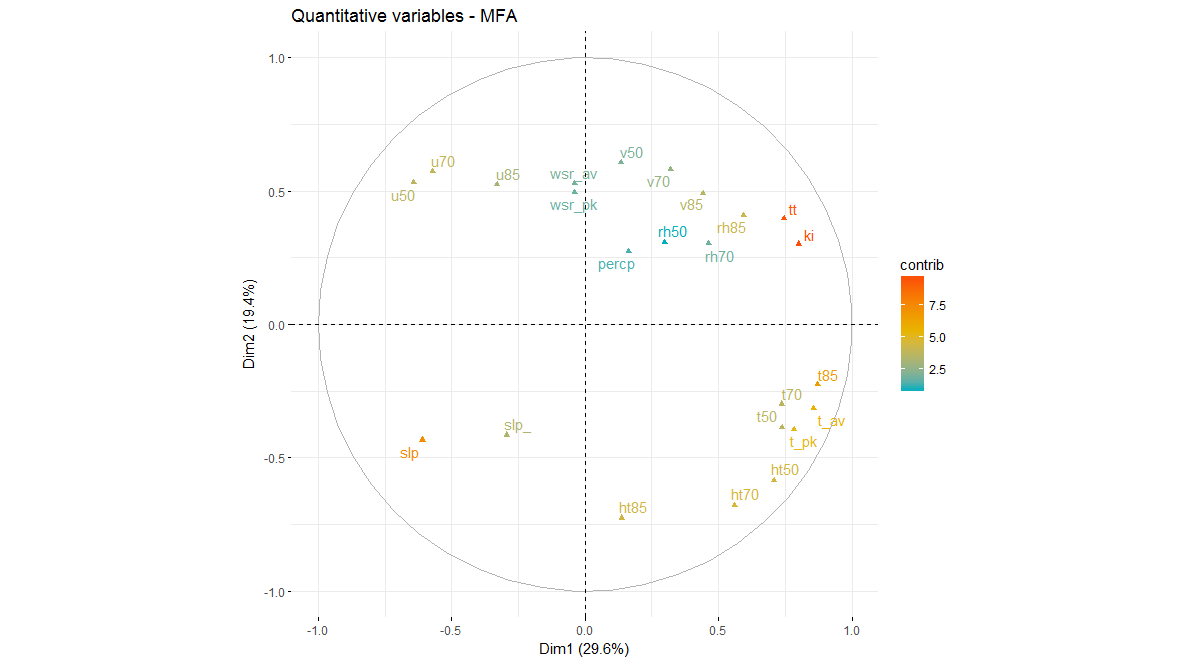
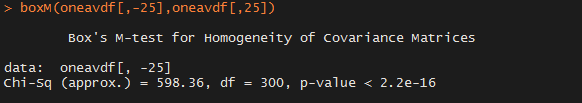


Figure Contribution of each variable to Factor 2



Figure



Figure

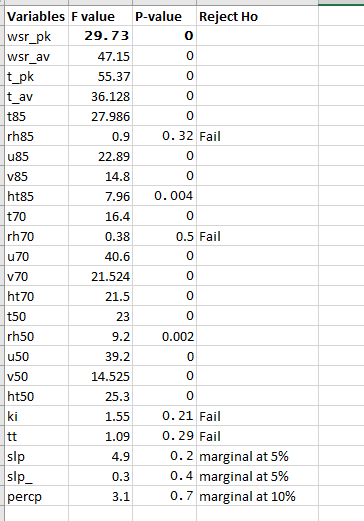


Figure Individual Anova Test

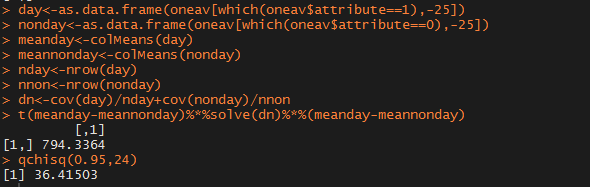


Figure Two sample Hotelling T2 test